Detecting innovations in a parsed corpus of learner English

Schneider, Gerold ; Gilquin, Gaëtanelle

Abstract: In research on L2 English, recent corpus-based studies indicate that some non- standard forms are shared by indigenized (ESL) and foreign (EFL) varieties of English, which challenges the idea of a clear dichotomy between innovation and error. We present a data-driven large-scale method to detect innovations, test it on verb + preposition structures (including phrasal verbs) and adjective + preposition structures, and describe similarities and differences between EFL and ESL. We use a dependency-parsed version of the International Corpus of Learner English to automatically extract potential innovations, defined as patterns of overuse compared to the British National Corpus as reference corpus. We measure overuse by means of collocation measures like O/E or T-score, and compare our results with similar results for ESL. In both quantitative and qualitative analyses, we detect similarities between the two varieties (e.g. discuss about) and dissimilarities (e.g. accuse for, only distinctive for EFL). We report more verb/adjective + preposition combinations than previous studies and discuss the roles of analogy and transfer.

DOI: https://doi.org/10.1075/bct.98.03sch

Detecting innovations in a parsed corpus of learner English

Gerold Schneider
University of Konstanz & University of Zurich
Gaëtanelle Gilquin
University of Louvain & FNRS

ABSTRACT

In research on L2 English, recent corpus-based studies indicate that some non-standard forms are shared by indigenized (ESL) and foreign (EFL) varieties of English, which challenges the idea of a clear dichotomy between innovation and error. We present a data-driven large-scale method to detect innovations, test it on verb + preposition structures (including phrasal verbs) and adjective + preposition structures, and describe similarities and differences between EFL and ESL. We use a dependency-parsed version of the International Corpus of Learner English to automatically extract potential innovations, defined as patterns of overuse compared to the British National Corpus as reference corpus. We measure overuse by means of collocation measures like O/E or T-score, and compare our results with similar results for ESL. In both quantitative and qualitative analyses, we detect similarities between the two varieties (e.g. *discuss about*) and dissimilarities (e.g. *accuse for*, only distinctive for EFL). We report more verb/adjective + preposition combinations than previous studies and discuss the roles of analogy and transfer.

1 Introduction

Since the era of Error Analysis, much focus in interlanguage studies has been on non-native-like features. Initially restricted to cases of misuse, the advent of learner corpus research has made it possible to identify cases of under- and/or overuse, which equally contribute to the non-nativeness of learner production (e.g. Nesselhauf 2005, Granger 2009, Salazar 2014: 180).

Recent theoretical and technological developments, however, have changed the way non-native features are considered and investigated. From a theoretical perspective, attempts have been made to bridge the paradigm gap that has long existed between research on learner language and on indigenized second-language varieties (cf. Mukherjee & Hundt 2011, Gilquin 2015a, Gut et al. 2015). Adopting an empirical approach, these studies have shown that learner English, or English as a Foreign Language (EFL), and indigenized varieties of English, or English as a Second Language (ESL), share certain non-standard features, be it in the domain of syntax (e.g. Edwards 2014), lexis/lexico-grammar (e.g. Nesselhauf 2009, Gilquin 2011, Gilquin & Granger 2011, Götz & Schilk 2011, Laporte 2012, Edwards & Laporte 2015) or phonology (e.g. Fuchs & Wunder 2015, Götz 2015). It has therefore become impossible to simply disregard any differences between EFL and ENL (English as a Native Language) as errors that should be eliminated, especially when they coincide with the "innovations" that are found in ESL.

From a technological perspective, it has become increasingly common to enrich learner corpora with different kinds of annotation (cf. van Rooy 2015), including syntactic annotation (e.g. Dickinson & Ragheb 2009, Rosén & Smedt 2010). This, in turn, has allowed for more sophisticated types of automatic data extraction, including extraction of L2 patterns (e.g. Schneider & Hundt 2009, Díaz-Negrillo et al. 2013, Ng et al. 2014).

In this paper, we take advantage of these theoretical and technological developments to examine non-native-like combinations of verb + prepositional phrase (PP) and adjective + PP. Starting from the assumption that not all non-native-like combinations are necessarily errors, we set out to identify potential instances of innovations in a corpus of learner English. The first steps of this identification are fully automatic, thanks to the syntactic parsing of the learner corpus and the comparison with a large parsed corpus of native English by means of collocation statistics. Such a corpus-driven approach is what distinguishes our study from most other studies that have sought to bridge the gap between EFL and ESL research (see above). It greatly facilitates the retrieval of phenomena that may be relatively rare in the data and would normally require large amounts of manual work (cf. Schneider & Zipp 2013).

In particular, we address the following research questions. First, can the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL and/or to ESL (RQ1)? Second, does the same method also allow us to detect which patterns of verb + PP and adjective + PP are more typical for EFL and which for ESL (RQ2)? Third, does the method give us the tools to find more patterns than have been previously described (RQ3)? Fourth, does the method give us the tools to distinguish between error and innovation (RQ4)?

The paper is structured as follows. In Section 2 we show that verb/adjective + PP constructions are an important characteristic of L2 and that using parsed data can lead to insightful observations. In Section 3, we present our data and our method using collocation statistics. In Section 4 we give quantitative results, comparing the triangular relationship between EFL, ESL and ENL, while in Section 5 we provide a

qualitative analysis. Section 6 addresses the question whether our method allows us to make a distinction between error and innovation, before the conclusion in Section 7.

2 Motivation

2.1 Verb + preposition and adjective + preposition combinations

In order to investigate differences between EFL, ESL and ENL use, one can in principle search for differences in linguistic patterns at any linguistic level: phonological, lexical, morphological, syntactic. The first of these is not available, given our selection of corpus data. According to Schneider (2004: 229), crucial differences between varieties occur at the level of lexico-grammar. It is the interaction between lexis and grammar that is open to variation, and it typically involves collocational preferences and verb complementation.

Collocational preferences can be captured by collocation measures, which we introduce in Section 3.1. Concerning complementation, we investigate combinations of verbs/adjectives and prepositions or verbal particles. Verb + PP combinations constitute an important and frequent (Cornell 1985) subgroup of verb complementation, and exhibit a high rate of innovation, both in ESL and EFL. In ESL, Indian English, for instance, presents a high degree of innovation in its use of prepositional verbs (Mukherjee & Hoffmann 2006); in EFL phrasal verbs represent "one of the most notoriously challenging aspects of English language instruction" (Gardner & Davies 2007: 339; see also Gilquin 2015b or Deshors to appear). New verb + PP combinations are a promising research object, as demonstrated by Nesselhauf (2009), who describes instances of combinations (e.g. *discuss about*, *enter into*, *request for*) which she found both in ESL and EFL varieties. The comparison between ESL and EFL also highlights the paradox that some of the "innovations" identified in ESL varieties coincide with those held up as common "errors" in EFL (cf. Gilquin to appear).

Prepositions have been shown to be difficult to acquire for non-native speakers of English, leading to avoidance, non-standard uses, etc. (see Gilquin & Granger 2011: 59-60). Investigations of selected prepositions and verbal particles, for example the preposition *into* (Gilquin & Granger 2011) or the particle *up* (Gilquin 2011), revealed interesting correlations between EFL, ESL and ENL use. Gilquin (2011) shows that both EFL and ESL speakers tend to overuse phrasal verbs in writing, while at the same time underusing them in speech, which indicates lacking ability to adapt to register conventions, although the degree differs, with ESL speakers being more sensitive to register variation. In order to address the question of how other prepositions and verbal particles pattern, the manual annotation work would be enormous. Fortunately, we can use automatic annotation, as shown in the following.

2.2 Syntactically parsed data

Corpus-based descriptions of ESL varieties (see Sand 2004, Schneider 2004 or Sedlatschek 2009) and EFL varieties (e.g. Nesselhauf 2005) have typically been

conducted on orthographic, i.e. not annotated, corpora. Automatic annotation has risks and benefits. It has the risk of errors adding to the noise of corpus imbalances, which is why we propose to use a semi-automatic approach, in which type-based candidates are presented to the user. For her investigation, Gilquin (2011) had to manually differentiate between *up* as a verbal particle and other uses, while we can now rely on automatically annotated data. Automatic annotation also offers the advantage that unrestricted amounts of data can be processed, which in comparison to Gilquin (2011, 2015a) allowed us to include the whole of the International Corpus of Learner English (ICLE), but also most components of the International Corpus of English (ICE) and the written part of the British National Corpus (BNC) (see Section 3.2 for a presentation of the corpora), and in addition made it possible to step up from selected particles/prepositions to all particles/prepositions, and all combinations they may have with their head verb, be they adjacent or not.

2.3 Innovation vs. error

Errors are traditionally associated with EFL, and innovations with ESL. However, the partial overlap between EFL and ESL non-standard features (see Section 1) means that the distinction between errors and innovations may have to be reconsidered. We start from the assumption that both errors and innovations may be found in either variety, and we seek to operationalize the distinction by objective means.

Van Rooy (2011: 189) points out that "[a] distinction between error and conventionalized innovation is essential to understanding if and how New Varieties of English develop new conventions". He suggests that the two key criteria for distinguishing innovations and errors are systematicity and acceptability.

Systematicity, which is required "to show that these variants are not mere random errors, but have found a place in emergent linguistic systems" (ibid.), is easily operationalized in our approach by means of collocation measures, and by discarding infrequent combinations (we discard hapax legomena). Acceptability is more difficult to operationalize. ICLE is not error-tagged, and there is no corpus-based way to find out if an innovative expression used by one EFL learner would be acceptable to other EFL learners. As far as the sparse data allows, we do check, however, if an expression is used by several learners, and if it is used by learners with different L1 backgrounds. The former may point to acceptability by a part of the community; the latter may point to a psycholinguistic base for an innovation, or to typologically related L1 backgrounds. Absence of the latter may indicate L1 transfer errors.

According to usage-based linguistics, acceptability typically follows from frequency, with a certain time lag. Frequency of co-occurrence is not only an effect of entrenchment, it is also often described as a contributor, as functional and cognitive linguists increasingly point out, e.g. Bybee (2007: 337). In practical terms, this entails that after a new combination (which is initially seen as an error) has occurred frequently enough and attains collocational status for some speakers, it has increasingly better chances to become accepted as an innovation.

Gut (2011: 120) notes that "[t]he labeling of a structure as an error (…) has an attitudinal and political rather than a linguistic basis". If this is the case, the systematicity-based continuum of chance co-occurrence to strong collocation, which can be directly measured by collocation statistics, may suffice as a first operationalization. We do not distinguish between innovation and error in Section 4, although the fact that we remove hapax legomena means that two very obvious types

of error, typos and single production errors, are excluded. We attempt to distinguish between innovation and error again in Sections 5 and 6 and give partial answers.

Gradient continua and attitudinal preferences can be captured by collocation statistics, which we introduce now.

3 Methods and Data

3.1 Method: Collocations and overuse

Schneider (2004: 229) mentions collocational differences as a feature of indigenized varieties of English. Collocations signify conventionalized use of linguistic expressions. Criteria include non-compositionality, non-substitutability, limited modifiability, non-literal translations and statistical co-occurrence. While only the last of these can be measured trivially in corpora, it has proven to be a surprisingly appropriate measure, both in terms of measuring collocation strength (e.g. Wulff 2008) and in approximating the psycholinguistic entrenchment which is behind collocations: Gries & Wulff (2005) and Gries & Wulff (2009) find strong correlations between collocation strengths and experimentally obtained sentence completions from advanced EFL learners, which means that collocation measures lend themselves as a model of listener expectations.

A wide array of frequency-based collocation statistics has been suggested, see e.g. Evert (2008) and Pecina (2009). We restrict our investigation to O/E and T-score. O/E (which literally means Observed divided by Expected) and its variant MI (Mutual Information) are information-theoretic measures (Shannon 1951) of the extent to which two words appear more often together (O=Observed) than expected (E) if all words were randomly distributed in the corpus (or inside the frame of a construction). O/E is defined as

$$\frac{O}{E} = \frac{p(x,y)}{p(x)\cdot p(y)} = \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\cdot\frac{f(y)}{N}} = \frac{f(x,y)\cdot N\cdot N}{f(x)\cdot f(y)\cdot N} = \frac{f(x,y)\cdot N}{f(x)\cdot f(y)}$$

where x is the first word, y is the second word, $p(x)$ is the independent probability of x, $f(x)$ is the frequency of x in the corpus, $p(x,y)$ is the joint probability of x and y occurring together, and N is the size of the corpus. If co-occurrence of x and y is due to chance, i.e. if there is no collocational force, then the independent probability of seeing both (*Expected*) and the joint probability of seeing the combination (*Observed*) are roughly equal.

In order to describe innovations in ESL and EFL, we need to find verb/adjective + PP combinations which (i) are conventionalized, i.e. frequent enough to reach collocation status, (ii) are collocations in the non-native corpora, and (iii) are not collocations, or much less so, in the native corpora. If we apply traditional collocation measures we fail to see point (iii). A successful measure for (iii) is the collocation ratio (Schneider & Zipp 2013): if $c_{L1}(x,y)$ is a collocation measure c for L1 of words x and y, then

Collocation ratio = $c_{L2}(x,y) / c_{L1}(x,y)$

The collocation ratio is a measure of overuse, of "overcollocability". Our suggested overuse statistics is an information-theoretic measure of surprise at seeing learner data when actually expecting native speaker data. For the collocation measure O/E, with c_{L2} as ICLE and c_{L1} as BNC, the ratio is defined as

$$O/E\ ratio = \frac{O/E(ICLE)}{O/E(BNC)} = \frac{\frac{O(ICLE)}{E(ICLE)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{ICLE}(R,w_1,w_2)\cdot N_{ICLE}}{O_{ICLE}(R,w_1)\cdot O_{ICLE}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2)\cdot N_{BNC}}{O_{BNC}(R,w_1)\cdot O_{BNC}(R,w_2)}}$$

where w_1=verb or adjective, w_2=preposition or verbal particle, R=syntactic relation expressing prepositional phrase attached to a verb, N=corpus size in words.

The O/E-ratio is itself an O/E measure, in which O=O/E(ICLE) and E=O/E(BNC), or in words: the observed value is the O/E measure as found in the application corpus ICLE, while we expected the O/E measure from the native speaker reference corpus BNC. O/E is an information theoretic measure of surprise: the interpretation of O/E-ratio is equally straightforward, it is also a measure of surprise.

The O/E measure has the tendency to over-represent rare events. The opposite characteristic has been attributed to the T-score measure. There are several answers to these two opposing characteristics. One is that as they are complementary, and if we thus apply both, we maximize recall. For the T-score collocation measure a formulation in terms of O and E (Evert 2008) is

$$T = \frac{O-E}{\sqrt{(O)}} \rightarrow T\ ratio = \frac{T(ICLE)}{T(BNC)} = \frac{\frac{O(ICLE)-E(ICLE)}{\sqrt{O(ICLE)}}}{\frac{O(BNC)-E(BNC)}{\sqrt{O(BNC)}}}$$

We also test and apply the T-ratio, but its statistical interpretation is more involved.

3.2 Data: Parsed EFL, ESL and ENL corpora

For the comparison of EFL, ESL and ENL, we use the following corpora. For EFL, we use the International Corpus of Learner English (ICLE; Granger et al. 2009). It is a corpus of learner English from university students with 16 different mother tongue backgrounds. It contains 3.7 million words from essays of higher intermediate to advanced learners of English.

For ESL, we use selected components of the International Corpus of English (ICE; Nelson et al. 2002). Each ICE component contains 1 million words of spoken and written text and has the same genre distribution. Among the 11 currently publicly available complete ICE components, 4 are from native language variants (GB, Canada, Ireland, New Zealand), while 7 contain ESL data, in which we are interested. We have excluded two components: ICE-East Africa, as it is made up of several subcomponents, and ICE Nigeria, as its spoken part contains no punctuation. We have kept all other ESL data, i.e. the following 5 components: ICE-Singapore, ICE-Philippines, ICE-Jamaica, ICE-India and ICE-Hong Kong.

For ENL, we use the written part of the British National Corpus (BNC; Aston & Burnard 1998). It contains 90 million words of written texts from a wide range of registers. We use it as a reference corpus of native British English.

We are aware that these corpora are not an ideal base for comparison: the mix of genres and the level of formality are different between the corpora: unedited student essays make up the entire ICLE but only small subsets of ICE, and have no counterpart in the BNC; they are also less formal than the written BNC, which consists largely of published material. This feature of the BNC, on the other hand, makes it suitable as a reference corpus of formal, high-level usage of British English. The ICE components which we use as an ESL reference have a much higher contingent of spoken language, which includes spontaneous, unedited usage. This characteristic is not only a disadvantage, but also an advantage when comparing the learner language represented in ICLE, which contains similarly spontaneous forms, many of which were not edited in the written essays, as the learners may not have been aware that they are infelicitous or incorrect. For these reasons, the ICE components are still a good alternative to the much larger GloWbE corpus (Davies & Fuchs 2015).

We use richly annotated corpus material: the corpora are annotated syntactically using the automatic dependency grammar parser Pro3Gres (Schneider 2008, Lehmann & Schneider 2012). An evaluation of the performance of the parser on ESL varieties is given and our approach is tested on selected phenomena in Schneider & Hundt (2009) and Schneider & Zipp (2013).

We do not distinguish between verbal particle and preposition, because often confusion between the two categories is at the core of the difference between the ENL use and the EFL or ESL use (e.g. *result in* vs. *result into*). For the same reason, we also include verb + PP combinations in which the PP is attached as an adjunct according to the automatic parser. We also include adjective + PP combinations, as they, too, have collocational status. For example, Benson et al. (2009) recognise adjective + preposition as an independent category in addition to verb + preposition (and noun + preposition, e.g. in nominalisations, which we have not included). Adjective + preposition combinations are often similarly difficult to acquire for learners of English.

4 Data-driven detection of verb/adjective + PP innovations/errors in EFL

In this section, we present and interpret our quantitative results. In the ranked lists that we show, we only give the top 10 to 30 entries, for space reasons. Our first operationalization of systematicity of innovations, which our algorithms (see Section 3.1) return and which we discuss, validate and interpret in the following, allows us to introduce a limited step of acceptability judgment by the authors, and a base for the qualitative analysis in Section 5.

4.1 Collocation ratio with O/E

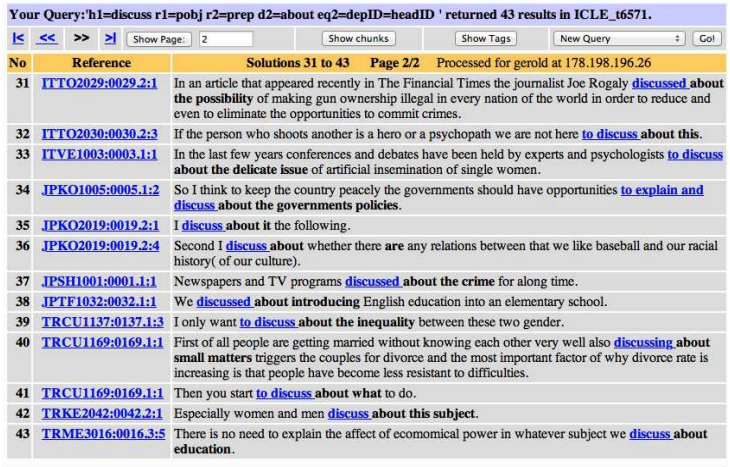
We first apply the O/E-ratio introduced in Section 3.1, using ICLE as application corpus and BNC as reference corpus. The top 30 candidates for EFL

innovations/errors are given in Table 1, sorted by decreasing O/E-ratio (first column). The second column contains the verb or adjective lemma, which is modified by the preposition or verbal particle given in column 3. Column 4 lists the frequency of the construction in ICLE. We have only excluded hapax legomena. Columns 5 and 6 give the collocation measure O/E for the application and reference corpora.

The last column is not output of our algorithm, but shows our comments and interpretation based on our inspection of the hits (see Figure 1, which lists the hits of line 24), in particular whether the type in this line is a learner innovation/error or not (for example because it is particularly frequent due to the essay topics, or a consistent parsing error). In uncertain cases we consulted dictionaries such as Benson et al. (2009). If our comment starts with "instead of" the hit is a true positive, i.e. the line represents a usage which is specific to learner English. The comment "CORPUS essay topic" means that this verb/adjective + preposition pair is overrepresented in ICLE because it appears very frequently due to the essay topics that are used in ICLE. *Handicap after*, for example, is overrepresented due to the essay topic "Discuss the pros and cons of abortion", where many students write that abortion should be allowed if a child would be *handicapped after birth*.

Table 1. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E-ratio

O/E ratio	VERB/ADJ.	PREP	F	O/E(ICLE)	O/E(BNC)	COMMENT
414.02	straight	out	2	1599.65	3.86	CORPUS essay topic
256.95	handicap	after	30	2211.46	8.61	CORPUS essay topic
201.30	responsible	of	19	23.31	0.12	instead of *responsible for*
150.95	worth	for	7	81.81	0.54	instead of *worth something*
144.47	view	upon	3	268.71	1.86	instead of *viewed on* (old-fashioned)
111.27	toss	about	2	505.05	4.54	
111.03	balance	from	2	47.87	0.43	
100.77	boil	by	2	45.97	0.46	
83.77	base	amongst	2	300.08	3.58	
77.10	attack	against	2	125.61	1.63	instead of *attack somebody*
72.87	alarm	of	2	92.95	1.28	
69.04	diverse	by	2	91.95	1.33	instead of *different according to*
65.18	exist	out	4	18.01	0.28	
53.54	design	before	2	304.28	5.68	
53.22	cool	down	4	6657.67	125.11	
50.78	bath	without	2	640.14	12.61	
50.31	sleep	around	13	420.93	8.37	
49.99	synonymous	to	2	26.10	0.52	instead of *synonymous with*
48.51	select	among	3	751.98	15.50	instead of *select from*
42.36	credit	for	2	233.73	5.52	
41.44	benefit	out	2	24.74	0.60	instead of *benefit from*
39.91	lower	than	4	198.58	4.98	
39.11	basic	for	2	58.43	1.49	
35.81	discuss	about	43	65.68	1.83	instead of *discuss something*
35.42	separate	between	4	189.54	5.35	instead of *distinguish between*
32.67	pour	onto	3	9928.44	303.87	
32.64	dependent	from	2	5.26	0.16	instead of *dependent on*
32.45	comment	by	2	22.19	0.68	
32.06	helpless	for	4	66.78	2.08	
31.47	stretch	beyond	4	6360.12	202.11	
30.22	understand	towards	2	54.88	1.82	instead of *understand sth.*



Your Query:'h1=discuss r1=pobj r2=prep d2=about eq2=depID=headID ' returned 43 results in ICLE_t6571.

|< << >> >| Show Page: 2 Show chunks Show Tags New Query Go!

No	Reference	Solutions 31 to 43 Page 2/2 Processed for gerold at 178.198.196.26
31	ITTO2029:0029.2:1	In an article that appeared recently in The Financial Times the journalist Joe Rogaly **discussed about the possibility** of making gun ownership illegal in every nation of the world in order to reduce and even to eliminate the opportunities to commit crimes.
32	ITTO2030:0030.2:3	If the person who shoots another is a hero or a psychopath we are not here **to discuss about this**.
33	ITVE1003:0003.1:1	In the last few years conferences and debates have been held by experts and psychologists **to discuss about the delicate issue** of artificial insemination of single women.
34	JPKO1005:0005.1:2	So I think to keep the country peacely the governments should have opportunities **to explain and discuss about the governments policies**.
35	JPKO2019:0019.2:1	I **discuss about it** the following.
36	JPKO2019:0019.2:4	Second I **discuss about** whether there **are** any relations between that we like baseball and our racial history(of our culture).
37	JPSH1001:0001.1:1	Newspapers and TV programs **discussed about the crime** for along time.
38	JPTF1032:0032.1:1	We **discussed about introducing** English education into an elementary school.
39	TRCU1137:0137.1:3	I only want **to discuss about the inequality** between these two gender.
40	TRCU1169:0169.1:1	First of all people are getting married without knowing each other very well also **discussing about small matters** triggers the couples for divorce and the most important factor of why divorce rate is increasing is that people have become less resistant to difficulties.
41	TRCU1169:0169.1:1	Then you start **to discuss about what** to do.
42	TRKE2042:0042.2:1	Especially women and men **discuss about this subject**.
43	TRME3016:0016.3:5	There is no need to explain the affect of ecomomical power in whatever subject we **discuss about education**.

BNC Dependency Bank 1.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

Figure 1. Hits for *discuss about* from ICLE, shown in Dependency Bank (Lehmann & Schneider 2012)

The last column of Table 1 thus shows that 12 of the top 30 candidates were indeed validated as EFL innovations/errors. In terms of the evaluation measure precision (e.g. Jurafsky & Martin 2009: 489)[1], this corresponds to 40% precision, which on the one hand may seem low, but on the other hand is sufficiently high, because manual filtering based on inspecting the hits is quite simple. We can easily increase precision by setting a filter on O/E(BNC) corresponding to the criterion that innovations/errors should not have high collocational status in the native variant. If we set a filter of O/E(BNC)<5, precision rises to above 50%, but at the trade-off of a cost in recall: for example, *select among* and *separate between* would not be returned. Equally, only including results which the automatic parser annotates as PP-arguments would increase precision, but lead to a large loss in recall. For example, *discuss about* and *attack against* are parsed as PP-adjuncts.

In Table 2, we use such a filter of O/E(BNC)<5, and in addition we take into consideration the fact that verb/adjective + preposition combinations which were not seen in the BNC may never appear there because they are unacceptable in native British English. We thus added a smoothing count of 0.5 (new fifth column) to types unseen in the BNC. We used a frequency threshold of f(ICLE)>3. As one can see in the last, again manually added column, 17 of the top 30 candidates (corresponding to 57% precision) are innovations/errors.

Table 2. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E ratio, with filter O/E(BNC)<5 and smoothing for events unseen in BNC

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	F(BNC)	O/E(ICLE)	O/E(BNC)	COMMENT
488.81	critical	towards	7	0.5	1511.26	3.09	instead of *critical to*
201.30	responsible	of	19	2	23.31	0.12	instead of *responsible for*
189.01	critical	against	4	0.5	370.22	1.96	instead of *critical to*
150.95	worth	for	7	1	81.81	0.54	instead of *worth something*
145.67	superior	than	22	0.5	434.65	2.98	instead of *superior to*
138.75	indulge	into	6	0.5	61.11	0.44	instead of *indulge in*
110.11	overcrowd	at	32	0.5	485.00	4.40	CORPUS essay topic
69.11	destructive	for	5	1	166.95	2.42	instead of *destructive to*
65.18	exist	out	4	2	18.01	0.28	
39.91	lower	than	4	2	198.58	4.98	
35.81	discuss	about	43	7	65.68	1.83	instead of *discuss something*
34.27	conscious	about	10	2	124.19	3.62	instead of *conscious of*
32.06	helpless	for	4	1	66.78	2.08	
31.55	possible	out	4	5	30.37	0.96	
30.60	recur	to	4	7	125.26	4.09	
29.94	dependent	of	8	4	19.34	0.65	instead of *dependent on*
24.63	belong	into	4	2	6.63	0.27	instead of *belong to*
23.59	renounce	to	9	3	108.40	4.60	
23.07	decide	over	7	13	102.14	4.43	CORPUS essay topic
21.96	inherent	to	9	13	78.29	3.56	
20.46	relate	with	49	76	32.98	1.61	instead of *relate to*
19.80	aware	about	4	1	5.94	0.30	instead of *aware of*
19.67	aspire	for	4	3	51.94	2.64	instead of *aspire to*
18.21	guilty	for	22	28	59.11	3.25	instead of *guilty of*
17.72	little	by	11	36	70.80	4.00	
17.67	produce	out	4	30	44.85	2.54	
17.19	accuse	for	8	19	18.33	1.07	instead of *accuse of*
15.39	interest	to	7	0.5	11.54	0.75	
15.01	specialize	on	4	4	40.24	2.68	
15.01	deal	about	4	2	3.91	0.26	instead of *deal with*

In Table 2, it is striking to see that the majority of true positives (12 out of 17) can be analysed as involving the use of a semantic, compositional preposition instead of a functional, idiomatic preposition, namely *critical towards*, *critical against*, *indulge into*, *destructive for*, *discuss about*, *conscious about*, *belong into*, *aware about*, *aspire for*, *guilty for*, *accuse for*, *deal about.*

4.2 Collocation ratio with T-score

We next apply T-ratio from Section 3.1, using ICLE as application corpus and BNC as reference corpus. The top 10 candidates for EFL innovations/errors are given in Table 3, sorted by decreasing T ratio (first column); all other columns are analogous. Figure 2 shows the hits of line 3.

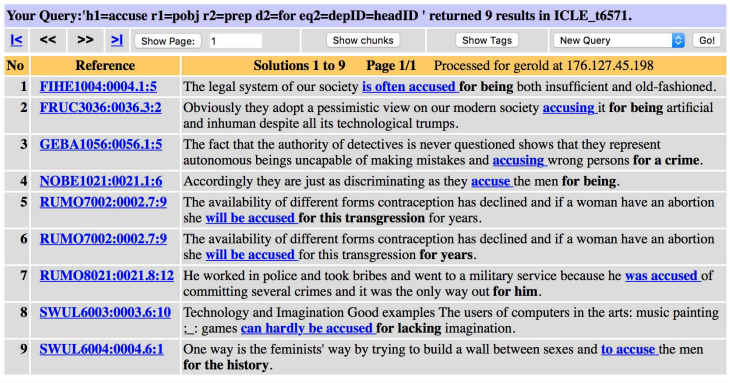
Table 3. Verb/adjective + preposition overuse in ICLE, sorted by decreasing T-ratio

T ratio	VERB/ADJ.	PREP	F	T(ICLE)	T(BNC)	COMMENT
5.9820	impose	to	10	5336.86	892.15	instead of *impose on*
3.5860	replace	to	3	1168.35	325.81	instead of *replaced by* (partly)
2.1133	accuse	for	8	5143.81	2433.98	instead of *accuse of*
2.0275	addict	on	4	3431.99	1692.68	instead of *addict to*
1.4296	better	than	87	17920.70	12535.47	
1.3929	alarm	of	2	2691.03	1932.01	instead of *alarm about*

[1] In words, precision measures how many hits are true positives; recall measures how many of all the true positives are found by the automatic system.

1.3322	handicap	after	30	10530.89	7905.03	CORPUS essay topic
1.2812	better	for	59	14564.98	11367.88	
1.2074	diverse	by	2	2690.71	2228.48	instead of *different according to*
1.1541	discuss	about	43	12421.43	10762.54	instead of *discuss sth.*

In terms of precision, 15 of the top 30 candidates with T-ratio are innovations/errors, which corresponds to 50% precision. We could increase precision by setting a filter on T(BNC) corresponding to the criterion that innovations/errors should not have high collocational status in the native variant. With a filter of, e.g., T(BNC)<5,000, precision rises to above 50%, but at the trade-off of a cost in recall (*discuss about* and *relate with*, for example, would not be returned).



Your Query:'h1=accuse r1=pobj r2=prep d2=for eq2=depID=headID ' returned 9 results in ICLE_t6571.

|< << >> >| Show Page: 1 Show chunks Show Tags New Query Go!

No	Reference	Solutions 1 to 9 Page 1/1 Processed for gerold at 176.127.45.198
1	FIHE1004:0004.1:5	The legal system of our society **is often accused for being** both insufficient and old-fashioned.
2	FRUC3036:0036.3:2	Obviously they adopt a pessimistic view on our modern society **accusing** it **for being** artificial and inhuman despite all its technological trumps.
3	GEBA1056:0056.1:5	The fact that the authority of detectives is never questioned shows that they represent autonomous beings uncapable of making mistakes and **accusing** wrong persons **for a crime**.
4	NOBE1021:0021.1:6	Accordingly they are just as discriminating as they **accuse** the men **for being**.
5	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she **will be accused for this transgression** for years.
6	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she **will be accused** for this transgression **for years**.
7	RUMO8021:0021.8:12	He worked in police and took bribes and went to a military service because he **was accused** of committing several crimes and it was the only way out **for him**.
8	SWUL6003:0003.6:10	Technology and Imagination Good examples The users of computers in the arts: music painting ;_: games **can hardly be accused for lacking** imagination.
9	SWUL6004:0004.6:1	One way is the feminists' way by trying to build a wall between sexes and **to accuse** the men **for the history**.

Dependency Bank 2.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

Figure 2. Hits for *accuse for* from ICLE, shown in Dependency Bank

4.3 Quantitative analysis of verb/adjective + PP combinations in ESL

We have so far detected EFL innovations/errors by our collocation-based approach. We can apply the same approach to ESL varieties; Schneider & Zipp (2013) have done so to describe innovations in ICE-Fiji and ICE-India.

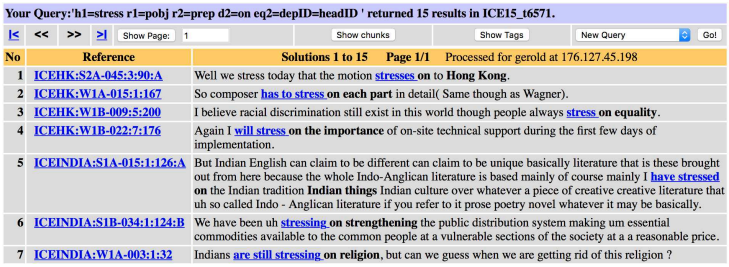
Any individual ESL variety could be analysed in the same fashion. Here, we use a collection of ESL varieties, the 5 ICE components described in Section 3.2, henceforth ICE-5 ESL. Using a collection of ESL corpora has the advantages that sparse data issues are reduced and that psycholinguistically based innovations, i.e. innovations not due to L1 transfer but due to general cognitive processes like analogy, are boosted. Analogy is seen as a key ability for language acquisition (Tomasello 2003) and generally in usage-based approaches to language (Bybee 2007). This mirrors our EFL approach of using all linguistic backgrounds in ICLE, but also has the disadvantage that innovations specific to one variety are likely to be overlooked.

Table 4 shows the top 22 candidates, again sorted by decreasing O/E-ratio. Some of the non-native-like combinations that we have seen in EFL also appear in ESL, for example *discuss about*. 6 of the top 22, for example *study about*, are innovations. As in ICLE, corpus buildup mismatches between application and reference corpus are responsible for some overused expressions (cf. "CORPUS"). The

tendency to transfer nominal subcategorisation patterns to the corresponding verb as in *emphasize* or *stress* (see Figure 3) may be a universal psycholinguistic mechanism (cf. Section 5).

Table 4. Verb/adjective + preposition overuse in ICE-5 ESL, sorted by decreasing O/E-ratio

O/E ratio	VERB/ADJ.	PREP	F	O/E(ICE-5 ESL)	O/E(BNC)	COMMENT
128.08	lower	than	12	637.31	4.98	
110.85	immerse	into	6	213.99	1.93	
55.27	canvass	before	31	2743.07	49.63	CORPUS: all from ICE-IND, legal term
54.04	preside	by	6	65.45	1.21	CORPUS: most from ICE-IND
50.31	play	inside	8	171.08	3.40	CORPUS: sports news
45.57	discuss	about	35	83.59	1.83	instead of *discuss sth.* / noun
35.95	understand	between	12	348.19	9.69	tagging error
28.70	elect	into	6	47.15	1.64	
26.90	emphasise	on	8	116.84	4.34	instead of noun
22.57	switch	over	12	292.55	12.96	instead of *switch to*
20.14	print	over	6	159.57	7.93	CORPUS: all from 1 ICE-JAM article
19.88	run	toward	11	515.50	25.94	CORPUS: sports news
19.76	study	about	14	26.05	1.32	instead of *study sth.* / noun
19.19	branch	into	6	505.80	26.36	
18.74	awaken	as	7	87.52	4.67	CORPUS: most from 1 ICE-IND article
18.15	coat	on	8	97.37	5.37	
16.73	better	than	80	1862.09	111.31	
16.67	sort	of	17	218.04	13.08	tagging error
14.92	accuse	before	6	123.02	8.24	
14.91	dress	on	8	39.61	2.66	
14.49	emphasize	on	9	69.18	4.78	instead of *emphasize sth.* /noun
13.35	stress	on	14	83.46	6.25	instead of *stress sth.* /noun



Your Query:'h1=stress r1=pobj r2=prep d2=on eq2=depID=headID ' returned 15 results in ICE15_t6571.

|< << >> >| Show Page: 1 Show chunks Show Tags New Query Go!

No	Reference	Solutions 1 to 15 Page 1/1 Processed for gerold at 176.127.45.198
1	ICEHK:S2A-045:3:90:A	Well we stress today that the motion **stresses on** to **Hong Kong**.
2	ICEHK:W1A-015:1:167	So composer **has to stress on each part** in detail(Same though as Wagner).
3	ICEHK:W1B-009:5:200	I believe racial discrimination still exist in this world though people always **stress on equality**.
4	ICEHK:W1B-022:7:176	Again I **will stress on the importance** of on-site technical support during the first few days of implementation.
5	ICEINDIA:S1A-015:1:126:A	But Indian English can claim to be different can claim to be unique basically literature that is these brought out from here because the whole Indo-Anglican literature is based mainly of course mainly I **have stressed on** the Indian tradition **Indian things** Indian culture over whatever a piece of creative creative literature that uh so called Indo - Anglican literature if you refer to it prose poetry novel whatever it may be basically.
6	ICEINDIA:S1B-034:1:124:B	We have been uh **stressing on strengthening** the public distribution system making um essential commodities available to the common people at a vulnerable sections of the society at a a reasonable price.
7	ICEINDIA:W1A-003:1:32	Indians **are still stressing on religion**, but can we guess when we are getting rid of this religion ?

Figure 3. Hits for *stress on* in ICE-5 ESL, shown in Dependency Bank

The generally smaller O/E-ratio in ESL as compared to EFL (Section 4.1) shows that ESL (represented by ICE-5 ESL) is closer to the BNC reference than is EFL (represented by ICLE). This is probably due to the following reasons. First, there are fewer innovations/errors in ESL than in EFL. Particularly errors, i.e. those choices which are not accepted by more experienced speakers of the same community, are less frequent in ESL. Second, the semantic similarities of the texts are probably less strong between individual ICE documents than between individual ICLE documents, which often have the same essay title. The fact that collecting many L1 backgrounds

glosses over many of the characteristics of an individual variety equally applies to ICE-5 ESL and to ICLE, and is therefore probably not a major reason for the large differences in O/E-ratio.

4.4 Quantitative analysis of verb/adjective + PP combinations in EFL vs. ESL

Until now we have compared EFL and ESL to a native British English reference. We can also compare EFL to an ESL reference corpus, indicating which innovations are more EFL-like. When using the same parameters as in Section 4.1 (Table 1), precision is quite low (6/30), indicating that EFL is closer to ESL than to the native reference corpus.

To boost precision, we ran a version of the innovation extraction method seen in Table 2, with particularly strict O/E(ICE 5 ESL)<2, counting unseen instances again as 0.5, aiming at a core set of typical verb/adjective + preposition innovations which only EFL speakers but not ESL speakers use. The top results are given in Table 5. 12 of the 21 top hits are true positives.

Looking at Table 5 reveals that noun-analogies (noun complementation patterns which are taken over to the verb) are very rare (only one, *assist to*) compared to ESL (Section 4.3, Table 4), and that the preposition *to* seems to be used too generically: 7 out of the 13 true positives involve *to*. There might be a trend to use *to* as a generic marker for indirect objects.

Table 5. Verb/adjective + preposition overuse in ICLE, sorted by decreasing O/E-ratio, using ICE-5 ESL as a reference corpus, with threshold O/E(ICE 5 ESL)<2, and smoothing for events unseen in ICE-5 ESL

O/E ratio	VERB/ADJ	PREP	F(ICLE)	F(ICE-5 ESL)	O/E(ICLE)	O/E(ICE-5 ESL)	COMMENT
35.97	equivalent	in	5	0.5	35.34	0.98	
34.19	assist	to	6	1	27.63	0.81	instead of *assist sth.*
25.68	accuse	for	8	0.5	18.33	0.71	instead of *accuse of*
22.29	wrong	at	6	0.5	24.38	1.09	
21.61	explain	from	8	0.5	16.03	0.74	
21.28	stay	like	5	0.5	13.53	0.64	
15.45	participate	to	8	1	8.46	0.55	instead of *participate in*
14.10	arise	by	6	0.5	12.14	0.86	instead of *due to/from*
12.60	employ	of	5	0.5	18.19	1.44	parsing error
11.35	benefit	to	13	1	10.49	0.92	instead of *be of benefit to*
9.10	impose	to	10	1	8.15	0.90	instead of *impose on*
8.06	oppose	in	6	0.5	5.05	0.63	
5.63	equal	for	9	0.5	4.22	0.75	instead of *equal to*
5.51	discuss	of	5	0.5	4.22	0.77	
5.40	remain	to	5	2	4.33	0.80	
5.34	necessary	with	6	0.5	6.70	1.25	instead of *necessary for*
5.08	keep	into	5	1	4.22	0.83	instead of *keep at*
5.05	reflect	to	5	1	5.12	1.01	instead of *reflect sth.*
4.95	confront	to	6	0.5	7.17	1.45	instead of *confront with*
4.93	discuss	for	13	2	6.13	1.24	
4.72	popular	to	6	0.5	4.84	1.03	instead of *popular for*

5 Qualitative analysis

We now examine the non-native-like verb/adjective + PP combinations found in EFL, using the method described above, and also briefly compare the results with those found for ESL. Our approach is more qualitative here, seeking to identify the processes that may have led to these combinations.

When compared to the native reference corpus BNC, some verb/adjective + PP combinations overused by learners are reported close to the top of the ranked lists by both O/E- and T-ratio, for example *basic for*, *discuss about*, *helpless for* or *relate with*. However, it also turns out that each measure brings up its own combinations. Interestingly, this includes the use of different prepositions with one and the same verb or adjective, for instance *independent from* (with the O/E-ratio) and *independent on* (with the T-ratio). This shows that neither of the two measures is sufficient in itself and that they should be combined with each other. Consequently, no distinction will be made between the two measures in the following qualitative analysis.

If we exclude typos, we can distinguish several types of major combinations. Some involve the use of a prepositional complement instead of a transitive use of the verb. Thus, instead of *discuss sth* some learners use the combination *discuss about* (1); instead of *consider sth* they use *consider about sth* (2); and instead of *phone sb* they use *phone to sb* (3).

1) First of all let's **discuss about** the goodness of having PC cafes. (ICLE:CNHK1224)
2) In this essay I am going to **consider about** the advantages and disadvantages of banning smoking in restaurants. (ICLE:CNHK1371)
3) He the boss **phoned to** his friends from Mafia and asked to get rid of his friends with whom he was bore to death. (ICLE:RUMO7025)

In other cases, it is the verb or the adjective that is inadequate. In example (4) the learner has used *insensible* instead of *insensitive*, and in example (5) *helpless* instead of *useless*.

4) One does not have to be a Marxist to understand what he meant: that religion was an escape from the hard everyday life making people ignorant and **insensible to** the wrongs that existed at that time. (ICLE:SWUL6001)
5) To conclude not all of qualifications people can get from universities are useful some of subjects are **helpless for** their future jobs (…). (ICLE:CNUK2015)

There are also many cases where a non-standard preposition is used, for example *concentrate to* (instead of *concentrate on*) and *intolerant to* (instead of *intolerant of*):

6) When the demand for these machines is big enough the production **concentrates to** certain areas and to certain people and the first step towards industrialism is then taken. (ICLE:FIJO3011)
7) As people usually get married at the young age they can be quite **intolerant to** any kind of disturbance in their new home. (ICLE:TRCU1169)

Very often, these non-native-like combinations have not been coined by chance, but seem to be the result of analogy, or more precisely "nativised semantico-structural analogy" (Mukherjee 2005, cited in Mukherjee & Hoffmann 2006). The basis of this analogy can be a word of the same family but with a different part-of-speech. The use

of *about* with the verb *discuss* (see example (1) above) could be related to the preposition that is used with the noun *discussion*. The same is true of *attack_V against* (cf. *attack_N against*), *be credited for* (cf. *credit_N for*) or *relate with* (cf. *relation with*), e.g. (8). In the case of *independent on*, cf. (9), it is the preposition of the positive form of the adjective, *dependent*, which is borrowed by the learners.

8) For example in the Gulf War the USA **attacked against** the Iraqis in order to prevent the price of petrol from going up. (ICLE:FIJO2003)
9) The first reason why childhood does not end when you become economically **independent on** your parents is that maturity is a mental condition which has nothing to do with money (…) (ICLE:ITVE2003)

In other combinations, the analogy is based on a synonym. Thus, the use of the preposition *between* after the verb *separate* (10) could be due to the use of the same preposition with the synonymous verb *distinguish*. *To be viewed upon as* (11) could be formed by analogy with *to be looked upon as*, *to arrive to* by analogy with *to get to*, and *afraid about* by analogy with *scared about*.

10) It looks like it can be hard to **separate between** what is reality and what is TV-entertainment. (ICLE:NOHO1037)
11) Women have always been **viewed upon as** the weaker part of the population that had to be led and helped by men. (ICLE:CZPR3005)

Finally, some combinations seem to be due to transfer from similar combinations in the learners' mother tongue, for example the use of *inherent to* by French-speaking learners (12), who have the combination *inhérent à* in their L1, where the preposition *à* corresponds to English *to*.

12) By reading ancient stories we realize that suffering is **inherent to** the human condition and we feel taken in a timeless feeling. (ICLE:FRUL1010)

When we compare these non-native-like combinations with those found for ESL, a number of similarities emerge. The combination *discuss about*, for example, is mentioned in the literature on Indian English (Mukherjee 2007), and also in a study that specifically compares EFL and ESL (Nesselhauf 2009). The results of Schneider & Zipp's (2013) study on ESL also partly overlap with our results, with combinations like *discuss about*, *benefit out of* and *aware about* being found in the two studies; compare (13) and (14). Similar phenomena are also attested in both analyses, like the use of a redundant particle, illustrated by *viewed upon as* (instead of *viewed as*) in (11) above or *listed down* in (15).

13) To sum it up I belive that the E. C will be a paradise for the middle-classes since they mostly have white-collar jobs often provided with a certain position they'll **benefit** most **out of** tax-deregulations etc. (ICLE:SWUL9013)
14) So they'll **benefit out of** the faculty teaching (ICE IND:S1A-064)
15) Adi Asenaca said an Asian Development Bank poverty participation survey **listed down** forms of poverty in the country and her ministry was following up on the recommendations. (ICE FJ:W2C 013)

At the same time, we also observe differences between EFL and ESL. If we consider the types of combinations found in the two varieties (cf. Section 4), it is striking that 7

of the 9 true positives in the ESL data (Table 4, Section 4.3) involve a preposition where Standard English would use an object-complement (e.g. *study sth* instead of *study about sth*). 5 of these (*discuss about*, *emphasise on*, *study about*, *emphasize on*, *stress on*) could be based on an analogy to noun usage. Our data suggests that analogy to the complementation patterns of nouns is particularly frequent among ESL speakers. In comparison, only 3 of the 12 true positive types showed a noun-analogy in EFL (Table 1, Section 4.1). On the other hand, innovations/errors involving the use of a semantic, compositional, often directional preposition instead of a functional, idiomatic preposition are slightly more common in EFL (12 out of 17, see Table 2) than in ESL (4 out of 9: *discuss about*, *study about*, *mention about*, *call as*). This indicates that ESL may prefer grammatical analogies, while EFL may overuse spatial and directional analogies.

6 Discussion

We now return to the discussion of error versus innovation. So far, the concept of innovation has mainly been limited to the description of native English and ESL in the literature. Yet, the presence of similar non-native-like combinations in EFL and ESL makes it difficult to maintain a sharp distinction between the two and treat these combinations as errors in the case of EFL and as innovations in the case of ESL. The results of our automatic detection of non-native-like combinations include instances that probably no one would want to consider linguistic innovations, for example misspellings. Others, however, should perhaps be treated on a par with ESL innovations.

In Section 4.3 (Table 4) we have learnt that analogy to the complementation patterns of nouns seems particularly frequent among ESL speakers. In Section 4.4 (Table 5) we have used an automatic method to detect those verb/adjective + preposition combinations that are considerably different in EFL and ESL. We can use the same method to detect those which are similar. For this purpose, we use the settings from Section 4.1 (Table 2), i.e. O/E(BNC)<5, f(ICLE)>3, smoothing for events unseen in BNC, but we only report those verb/adjective + preposition combinations whose O/E from ICLE is not very different from the one in ICE-5 ESL. As a threshold we set that O/E(ICLE) is maximally 3 times larger than O/E(ICE-5 ESL) or vice versa. Results are given in Table 6. These are verb/adjective + preposition combinations which, according to our data, are shared between EFL and ESL. As they exist independently in both, with similar O/E-ratios, we hypothesize that they are more likely to be based on psycholinguistic trends than on L1 transfer or acquisition processes.

Table 6. Results from Table 2, filtered for similar O/E in ICLE and ICE-5 ESL

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	O/E (ICLE)	O/E (ICE-5 ESL)	O/E (BNC)	COMMENT
145.67	superior	than	22	434.6	565.61	2.98	instead of *superior to*
138.75	indulge	into	6	61.11	28.10	0.44	instead of *indulge in*
35.81	discuss	about	43	65.68	83.59	1.83	instead of *discuss sth*.
34.27	conscious	about	10	124.1	78.30	3.62	instead of *conscious of*
19.67	aspire	for	4	51.94	31.93	2.64	instead of *aspire to*
17.72	little	by	11	70.80	38.50	4.00	

15.39	interest	to	7	11.54	6.08	0.75	
14.29	point	by	6	13.23	5.57	0.93	
13.49	commensurate	to	4	22.37	49.29	1.66	
13.24	interest	for	26	63.97	41.70	4.83	
12.94	speak	over	5	33.16	13.06	2.56	
10.65	own	to	8	23.20	8.80	2.18	instead of *owing to* (partly)
10.28	watch	than	4	17.52	18.76	1.70	
9.75	capable	in	5	2.83	2.97	0.29	instead of *capable of/to*
9.10	deprive	from	10	18.64	12.64	2.05	
8.84	study	about	8	11.66	26.05	1.32	instead of *study sth*.
8.62	charge	of	4	30.98	11.88	3.59	instead of *change sth*/noun
7.86	shut	to	7	36.53	27.73	4.65	
7.28	face	to	35	19.64	7.86	2.70	instead of *face sth*.
7.24	state	about	4	25.04	11.77	3.46	
6.81	invest	to	5	5.44	2.93	0.80	instead of *invest in*
6.66	speed	in	5	33.13	27.33	4.98	
6.65	waste	for	8	24.28	18.73	3.65	
6.52	reward	to	6	18.07	24.65	2.77	
6.37	associate	to	4	3.89	3.29	0.61	instead of *associate with*
6.36	strike	to	6	16.48	6.16	2.59	
6.02	know	over	4	16.60	9.30	2.76	
5.95	afford	with	4	18.63	33.91	3.13	
5.89	steal	to	6	9.39	3.21	1.59	instead of *steal from*
5.88	sum	in	4	22.32	30.50	3.80	
5.51	influence	on	15	15.21	6.40	2.76	instead of noun(partly)
5.30	depend	from	9	4.84	1.76	0.91	instead of *depend on*
5.19	search	from	5	15.06	7.52	2.90	instead of *search on*

Among the combinations which can be treated on a par, it is important to distinguish between combinations that seem to be the result of L1 influence and those that seem to be the result of cognitive operations such as analogy. The latter, which we have called psycholinguistically based innovations, are probably more likely to be recognized as innovations than the former (L1 transfer innovations). Table 6 includes particularly many types that can be due to analogy, as we show in Table 7, which filters by those types that are found in speakers from several L1 backgrounds, (penultimate column). In the last column, we suggest a possible analogy. For example (as discussed in Table 2: a semantic preposition replaces a functional one), in *indulge into* the preposition iconically reduplicates a directionality instigated by the verb; in *aspire for* the subcategorisation frame is derived from the corresponding nominalisation. In future research, we want to test if ESL (and ENL) speakers are more willing to accept unusual patterns based on analogy than other patterns.

Table 7. Possible analogy interpretations of innovations which are common to EFL and ESL

O/E ratio	VERB/ADJ.	PREP	COMMENT	# L1 BACKG.	POSSIBLE ANALOGY
145.67	superior	than	instead of *superior to*	8	better than
138.75	indulge	into	instead of *indulge in*	4	iconic
35.81	discuss	about	instead of *discuss sth*.	7	discussion (noun)
34.27	conscious	about	instead of *conscious of*	6	
19.67	aspire	for	instead of *aspire to*	3	aspiration (noun)
10.65	own	to	instead of *owing to* (partly)	7	
9.75	capable	in	instead of *capable of/to*	4	diligent in
8.84	study	about	instead of *study sth*.	4	

8.62	charge	of	instead of noun	3	noun
7.28	face	to	instead of *face sth*.	>9	face *up to* w/o *up*
6.81	invest	to	instead of *invest in*	2	devote to
6.37	associate	to	instead of *associate with*	4	relate to
5.89	steal	to	instead of *steal from* (partly)	6	
5.51	influence	on	instead of noun(partly)	4	noun
5.30	depend	from	instead of *depend on*	3	iconic
5.19	search	from	instead of *search on*	2	

The purpose of one's approach should also be taken into account when trying to identify innovations. For descriptive purposes, one might be more inclined to recognize the learner's right to be creative, and hence the existence of linguistic innovations, whereas for pedagogical purposes teachers will teach native-like combinations and reject most non-native-like combinations – and perhaps rightly so. Finally, the setting is important too. In an EFL setting, which focuses on competence, non-native-like combinations are less likely to be accepted as innovations than in an English as a Lingua Franca setting, where communication takes precedence over competence.

7 Conclusion

We have described innovations in verb/adjective + preposition combinations (including phrasal verbs) in learner English, using ICLE as application corpus and BNC as reference corpus. We have applied overuse statistics like O/E and T-score, known from collocation analysis to detect and describe errors and innovations in learner English. Overuse statistics are an information-theoretic measure of surprise at seeing learner data when actually expecting native speaker data. We have given a first evaluation of the precision of our method, which allows us to answer the first part of RQ1 (can the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL and/or to ESL?) positively: the patterns of overuse which we observe using collocation statistics deliver combinations that are specific to EFL language. Our method, which we call collocation ratio, is corpus-driven and as far as we are aware reports more combinations than have previously been described (it should be borne in mind that for space reasons we could only show the top entries of considerably longer lists). Using more and larger EFL and ESL corpora would likely deliver further patterns. We can thus also answer RQ3 (does the method give us the tools to find more patterns than have been previously described?) positively..

We have applied the same method to ESL varieties using selected components from ICE and have provided a first evaluation, which allows us to answer the second part of RQ1 positively: the patterns of overuse which we observe using collocation statistics also deliver combinations that are specific to ESL language. In order to assess differences between EFL and ESL, we have compared EFL data against ESL data as a reference. This delivers combinations which are likely to be seen as unacceptable by ESL speakers, and are thus candidates for errors. Concerning RQ2 (does the same method also allow us to detect which patterns of verb + PP and adjective + PP are more typical for EFL and which for ESL?), we thus give a tentatively positive answer. Our data suggests that analogy to the complementation patterns of nouns is particularly frequent among ESL speakers, while EFL speakers

tend to overuse the preposition *to*. The use of compositional, semantic prepositions instead of idiomatic, functional ones (e.g. *indulge into*, *discuss about*) seems to be a shared pattern.

In order to assess similarities between EFL and ESL, we have further performed a qualitative analysis, and we have also reported which verb/adjective + preposition innovations in ESL attain similar O/E ratios in EFL. The approach comparing O/E ratios delivers combinations which are likely to be seen as acceptable by ESL speakers, and are thus candidates for innovations. The instances found in this way all occur in a variety of L1 backgrounds, which increases the probability that they are not caused by L1 transfer, but are based on more psycholinguistic mechanisms such as processes of analogy (e.g. the subcategorisation frame is derived from the corresponding nominalisation) or iconicity (e.g. the preposition iconically reduplicates a directionality instigated by the verb). In the qualitative step of our analysis, we have discussed relevant examples and performed a manual classification of the combinations. We infer that neither O/E nor T-score are sufficient on their own, as each brings up results that the other misses, and that they thus need to be combined to increase recall.

Concerning RQ4 (does the method give us the tools to distinguish between error and innovation?), we have partly narrowed down the candidates by excluding hapax legomena, by restricting innovations to combinations that are found in both EFL and ESL, and that are used by speakers of several L1 backgrounds. We have also singled out cases that can be explained by analogy. However, the results obtained cannot be evaluated, unlike in the other RQs. On the one hand this means that we can only give a speculative answer to RQ4, on the other hand it means that we are treading on new scientific ground by presenting lists of shared verb/adjective + PP combinations to the research community.

Our method thus offers a powerful means of automatically extracting from corpora a large number of patterns distinctive for EFL and/or ESL, and gives some clues as to the status of these patterns (errors or innovations). It therefore contributes to the recent efforts to bridge the paradigm gap between EFL and ESL, by providing new techniques that facilitate the analysis and should make it possible to collect further evidence for the link between the two varieties.

8 References

Aston, G. & Burnard, L. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Benson, M., Benson, E. & Ilson, R. 2009. *The BBI Combinatory Dictionary of English*. 3rd edition. Amsterdam and Philadelphia: John Benjamins.

Bybee, J. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Cornell, A. 1985. "Realistic goals in teaching and learning phrasal verbs", *International Review of Applied Linguistics in Language Teaching (IRAL)* 23(4), 269-280.

Davies, M. & Fuchs, R. 2015. "Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE)", *English World-Wide* 36(1), 1-28.

Deshors, S. C. To appear. "Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations", *International Journal of Learner Corpus Research* 2(1).

Díaz-Negrillo, A., Ballier, N. & Thompson, P. (Eds.). 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. Studies in Corpus Linguistics 59. Amsterdam and Philadelphia: John Benjamins.

Dickinson, M. & Ragheb, M. 2009. "Dependency annotation for learner corpora". In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT)*. Milan, Italy.

Edwards, A. 2014. "The EFL-ESL continuum and the case of the Netherlands: A comparative analysis of the progressive aspect", *World Englishes* 33, 173-194.

Edwards, A. & Laporte, S. 2015. "Outer and expanding circle Englishes. The competing roles of norm orientation and proficiency levels", *English World-Wide* 36(2), 135-169.

Evert, S. 2008. "Corpora and collocations". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: de Gruyter, 1212-1248.

Fuchs, R. & Wunder, E.-M. 2015. "A sonority-based account of speech rhythm in Chinese learners of English". In U. Gut, R. Fuchs & E.-M. Wunder (Eds.), *Universal or Diverse Paths to English Phonology? Bridging the Gap between Research on Phonological Acquisition of English as a Second, Third or Foreign Language*. Berlin: de Gruyter, 165-184.

Gardner, D. & Davies, M. 2007. "Pointing out frequent phrasal verbs: A corpus-based analysis", *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 41(2), 339-359.

Gilquin, G. 2011. "Corpus linguistics to bridge the gap between World Englishes and Learner Englishes". In L. Ruiz Miyares & M. R. Álvarez Silva (Eds.), *Comunicación social en el siglo XXI, Vol. II*. Santiago de Cuba: Centro de Lingüística Aplicada, 638–642.

Gilquin, G. 2015a. "At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared", *English World-Wide* 36(1), 91-124.

Gilquin, G. 2015b. "The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach", *Corpus Linguistics and Linguistic Theory* 11(1), 51-88.

Gilquin, G. To appear. "Applied cognitive linguistics and second/foreign language varieties: Towards an explanatory account". In E. Tribushinina, J. Evers-Vermeul & L. Rasier (Eds.), *Usage-based Approaches to Language Acquisition and Language Teaching*. Berlin: de Gruyter.

Gilquin, G. & Granger, S. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam and Philadelphia: John Benjamins, 55-78.

Götz, S. 2015. "Fluency in ENL, ESL and EFL: A corpus-based pilot study". In *Proceedings of Disfluency in spontaneous speech, DISS 2015*. Glasgow, UK. Available at: http://disfluency.org/DiSS_2015/Programme_files/Goetz-DISS2015.pdf (accessed April 2016).

Götz, S. & Schilk, M. 2011. "Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners". In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam and Philadelphia: John Benjamins, 79-100.

Granger, S. 2009. “Prefabricated patterns in advanced EFL writing: Collocations and formulae”. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 185-204.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, S. T. & Wulff, S. 2005. “Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora”, *Annual Review of Cognitive Linguistics* 3, 182-200.

Gries, S. T. & Wulff, S. 2009. “Psycholinguistic and corpus linguistic evidence for L2 constructions”, *Annual Review of Cognitive Linguistics* 7, 163-186.

Gut, U. 2011. “Studying structural innovations in New English varieties”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam and Philadelphia: John Benjamins, 100-124.

Gut, U., Fuchs, R. & Wunder, E.-M. (Eds.). 2015. *Universal or Diverse Paths to English Phonology*. Berlin: de Gruyter.

Jurafsky, D. & Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Second edition. Upper Saddle River, NJ: Prentice Hall.

Laporte, S. 2012. “Mind the gap! Bridge between World Englishes and Learner Englishes in the making”, *English Text Construction* 5, 265-292.

Lehmann, H. M. & Schneider, G. 2011. “A large-scale investigation of verb-attached prepositional phrases”. In S. Hoffmann, P. Rayson & G. Leech (Eds.), *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Varieng, Helsinki. Available at: http://www.helsinki.fi/varieng/series/volumes/06/lehmann_schneider/ (accessed April 2016).

Lehmann, H. M. & Schneider, G. 2012. “Dependency Bank”. In *Proceedings of LREC 2012 Workshop on Challenges in the Management of Large Corpora*, 23-28.

Mukherjee, J. 2005. “All mine, mine alone…”. Emerging local norms in Indian English lexico-grammar. Paper presented at the University of Zurich.

Mukherjee, J. 2007. “Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium”, *Journal of English Linguistics* 35(2), 157-187.

Mukherjee, J. & Hoffmann, S. 2006. “Describing verb-complementational profiles of New Englishes: A pilot study of Indian English”, *English World-Wide* 27(2), 147-173.

Mukherjee, J. & Hundt, M. 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam and Philadelphia: John Benjamins.

Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World: G29. Amsterdam and Philadelphia: John Benjamins.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam and Philadelphia: John Benjamins.

Nesselhauf, N. 2009. “Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties”, *English World-Wide* 30(1), 1-25.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. & Bryant, C. (Eds.). 2014. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Baltimore, Maryland, June.

Pecina, P. 2009. *Lexical Association Measures: Collocation Extraction. Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Charles University in Prague.

Rosén, V. & Smedt, K. D. 2010. “Syntactic annotation of learner corpora”. In Johansen, H., Golden, A., Hagen, J. E. & Helland, A.-K. (Eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag* [Systematic, Varied, but not Arbitrary. Anthology about Norwegian as a Second Language on the Occasion of Kari Tenfjord’s 60th Birthday]. Oslo: Novus forlag, 120-132.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. 2001. Multi-word expressions: A pain in the neck for NLP. Technical Report LinGO Working Paper No. 2001-03, Stanford University, CA.

Salazar, D. 2014. *Lexical Bundles in Native and Non-native Scientific Writing*. Amsterdam and Philadelphia: John Benjamins.

Sand, A. 2004. “Shared morpho-syntactic features in contact varieties of English: Article use”, *World Englishes* 23(2), 281-98.

Schneider, E. W. 2004. “How to trace structural nativization: Particle verbs in world Englishes”, *World Englishes* 23(2), 227-249.

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. PhD Thesis. Institute of Computational Linguistics, University of Zurich.

Schneider, G. & Hundt, M. 2009. “Using a parser as a heuristic tool for the description of New Englishes”. In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Schneider, G. & Zipp, L. 2013. “Discovering new verb-preposition combinations in New Englishes”. *Studies in Variation, Contacts and Change in English* 13. Available at: http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp (accessed April 2016).

Sedlatschek, A. 2009. *Contemporary Indian English: Variation and Change*. Amsterdam and Philadelphia: John Benjamins.

Shannon, C. 1951. “Prediction and entropy of printed English”, *The Bell System Technical Journal* 30, 50-64.

Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Van Rooy, B. 2011. “A principled distinction between error and conventionalized innovation in African Englishes”. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging the Paradigm Gap*. Amsterdam and Philadelphia: John Benjamins, 189-207.

Van Rooy, B. 2015. “Annotating learner corpora”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 79-105.

Wulff, S. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. London: Continuum.